ENABLING NATIONAL-SCALE GENOMICS ON THE CLOUD

Gaurav Kaushik, PhD
ISMB AKES Workshop
July 9th, 2016
TCGA IS A TREMENDOUS GIFT TO THE COMMUNITY

More than 11,000 cases representing 33 cancer types
TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY …

- Primary Tumor
  - Metastatic

- Blood Derived Normal
  - Solid Tissue Normal

- Genomic
  - Proteomic

- Transcriptomic
  - Epigenomic

multiple Samples per Case

multiple Analyses per Sample
... WITH FAR REACHING IMPACT.
As the amount and diversity of data increases, it becomes more difficult to learn from it.
ACCESSING DATA AT PETABYTE-SCALE

Repository

2.5 PB

...Downloading...

...Download 20,000,000 Gigabits of TCGA data...

...at 10 Gbps...

...will take over 23 days
THE COST OF BIG DATA

$2M/year in storage costs

Data is locked away and replicated unnecessarily across many institutions

Requires significant computational resources

Collaborating in real-time and sharing reproducible results is challenging.
3 YEARS IN THE MAKING...

April 2013: Recognizing these challenges, Dr. Harold Varmus & colleagues issue letter proposing creation of public “cancer knowledge clouds” and seeking input from the research community on data storage and compute challenges.
3 YEARS IN THE MAKING...

June 2013: The Cancer Genomics Cloud Pilot concept, presented by Dr. George Komatsoulis receives unanimous approval at a joint meeting of the NCI Board of Scientific Advisors and the National Cancer Advisory Board.
August 2013: Community feedback regarding capabilities and critical use cases collected via IdeaScale site & Sources Sought notice
September 2014: The Broad Institute, Institute for Systems Biology and Seven Bridges awarded two year contracts to build pilot systems.
THE CANCER GENOMICS CLOUD
GUIDING PRINCIPLES
GUIDING PRINCIPLES

Great science happens in teams

Making data available isn’t enough to make it usable

Reproducibility shouldn’t be hard

The impact of TCGA is extended by new data & tools
COLLABORATE EASILY

Create a project workspace to organize collaborators, data, and analyses.
COLLABORATE EASILY

Fine grained permissions allow control over who can see, use, and modify project assets.
TCGA Controlled data projects access limited to only Authorized users.

COMPLIANT COLLABORATION
PUBLIC PROJECTS

Cancer Cell Line Encyclopedia Dataset and Public Project

The Cancer Cell Line Encyclopedia (CCLE) is made possible through a collaboration between the Broad Institute, the Novartis Institutes for Biomedical Research, and the Genomics Institute of the Novartis Research Foundation to perform detailed genetic and pharmacologic characterization of a large number of human cancer models.

This project contains Open Access sequencing data (in the form of reads aligned to the hg19 broad variant genome reference) for nearly 1000 cancer cell line samples, as available from GitHub on May 11, 2016. Use of these data does not require any special access or authorization status or data storage costs, making them ideal for interrogating the genomic landscape of cancer cell lines, testing new analysis methods, or getting to know the Seven Bridges Platform.
GUIDING PRINCIPLES

Great science happens in teams

Making data available isn’t enough to make it usable

Reproducibility shouldn’t be hard

The impact of TCGA is extended by new data & tools
KNOW THE DATA BETTER

An interactive, living overview of TCGA
KNOW THE DATA BETTER

An interactive, living overview of TCGA
GET TCGA DATA FASTER

Query TCGA using >140 properties about cases, samples, files & more...
DATASETS API
## Public Reference Files

<table>
<thead>
<tr>
<th>File name</th>
<th>Size</th>
<th>File extension</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ucsc.hg19.fasta.fai</td>
<td>3.5 KB</td>
<td>FAI</td>
<td>Index file for ucsc.hg19.fasta. It is the same as human_g19.hg19.fai</td>
</tr>
<tr>
<td>ucsc.hg19.fasta</td>
<td>3.0 GB</td>
<td>FASTA</td>
<td>Human reference sequence UCSC hg19 assembly. It is the same as human_g19.hg19.fai</td>
</tr>
<tr>
<td>snpEff_v3_6_GRCh37_75.zip</td>
<td>168.9 MB</td>
<td>ZIP</td>
<td>SnpEff database file. Should be used with snpEff version 3.6. source: http://source_snpEff/</td>
</tr>
</tbody>
</table>
EXPLORE PROCESSED DATA WITH PRE-BUILT OR CUSTOM VIZ
IMMEDIATELY RUN AN ANALYSIS

Public apps for your data analysis

Browse our publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

Search or Explore all apps

RNA-Seq Alignment - STAR

Toolkit: STAR 2.4.2a

Alignment to a reference genome and transcriptome presents the first step of RNA-Seq analysis. This pipeline uses STAR, an ultrafast RNA-seq aligner capable of mapping full-length RNA sequences and detecting de novo canonical junctions, non-canonical exons, and...
Deep versioning makes it easy to track the history of workflow development and allows you to revert to a different version.
GUIDING PRINCIPLES

Great science happens in teams

Making data available isn’t enough to make it usable

Reproducibility shouldn’t be hard

The impact of TCGA is extended by new data & tools
EACH TASK IS REPLICABLE & REMEMBERABLE
DETAILED TASK LOGS AND STATISTICS ALLOW OPTIMIZATION
ROBUST PUBLIC API

• >50 calls
• Improved response
• Complete docs and training resources
• Highlights:
  • Task stats
  • Query by metadata
  • Easy batch execution model
LAUNCH THOUSANDS OF TASKS
GUIDING PRINCIPLES

Great science happens in teams

Making data available isn’t enough to make it usable

Reproducibility shouldn’t be hard

The impact of TCGA is extended by new data & tools
4 WAYS TO ADD DATA

Graphical uploader

Command Line

FTP / HTTP

API
**ANNOTATE YOUR DATA**

Metadata makes data findable and makes it easier to query and integrate.
EASILY FILTER FILES BY METADATA/PROPERTIES

<table>
<thead>
<tr>
<th>File name</th>
<th>Experimental strategy</th>
<th>Task ID</th>
<th>Created on</th>
<th>Size</th>
<th>Reference genome</th>
<th>Sample ID</th>
<th>Sample type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0732</td>
<td>WXS</td>
<td>ea1f1a0d0...</td>
<td>Feb. 9, 2016</td>
<td>5.3 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0732</td>
<td>WXS</td>
<td>86feb20a...</td>
<td>Feb. 9, 2016</td>
<td>8.6 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0733-08.4.sam</td>
<td>WXS</td>
<td>7f09279b...</td>
<td>Feb. 9, 2016</td>
<td>9.1 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0733-08.2.sam</td>
<td>WXS</td>
<td>8266d037...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0734-08.2.sam</td>
<td>WXS</td>
<td>9666a8a9...</td>
<td>Feb. 9, 2016</td>
<td>9.1 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0734-08.2.sam</td>
<td>WXS</td>
<td>6546c39e...</td>
<td>Feb. 9, 2016</td>
<td>9.1 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0735-08.2.sam</td>
<td>WXS</td>
<td>11175de6...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0735-08.2.sam</td>
<td>WXS</td>
<td>5b46c39e...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0736-08.2.sam</td>
<td>WXS</td>
<td>1b46c39e...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0736-08.2.sam</td>
<td>WXS</td>
<td>05771f22...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0737-08.2.sam</td>
<td>WXS</td>
<td>06771f22...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0737-08.2.sam</td>
<td>WXS</td>
<td>07771f22...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-03A-01W-0738-08.2.sam</td>
<td>WXS</td>
<td>08771f22...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-03A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
<tr>
<td>C17.TCGA-AB-2895-11A-01W-0738-08.2.sam</td>
<td>WXS</td>
<td>09771f22...</td>
<td>Feb. 9, 2016</td>
<td>11.2 MB</td>
<td>HG19_Broad_variant</td>
<td>TCGA-AB-2895-11A Primary Blood Derived Cancer - Peripheral Blood</td>
<td></td>
</tr>
</tbody>
</table>
AS THE AMOUNT OF DATA HAS GROWN, SO TOO HAS THE NUMBER OF TOOLS AVAILABLE TO ANALYZE IT.

11,000+ -omics data analysis tools*
  (each with many versions)

50+ used in a single TCGA marker paper

*omictools.com
DOCKER + COMMON
WORKFLOW LANGUAGE

COMMON
WORKFLOW LANGUAGE

DOCKER + COMMON
WORKFLOW LANGUAGE
COMMON WORKFLOW LANGUAGE

A COMMUNITY-DEVELOPED SPECIFICATION FOR PORTABLE ANALYSES
COMMON WORKFLOW LANGUAGE

20+ code contributors from 15+ organizations

170+ members of the user group

Bi-monthly meetings to discuss development
Everything you need to run an application contained in one, lightweight text file
RABIX SDK FOR CWL

A open-source development kit for composing and executing CWL applications in a portable, reproducible, and scalable fashion

Local CWL executor project (beta) at github.com/rabix/bunny
GET INVOLVED

Register at genomicscloud.org

Get an account with $100 in comp. credits

Get started at docs.cancergenomicscloud.org
REQUEST ADDITIONAL FUNDS

Receive $500 + $1000 by giving feedback

Request up to $10K for your project

(learn more at genomicscloud.org)
THANK YOU!